The physical mandate for belief-goal psychology

Article (Published Version)

Minds and Machines
https://doi.org/10.1007/s11023-020-09515-w

The Physical Mandate for Belief-Goal Psychology

Simon McGregor[1] · **Ron Chrisley**[1]

Received: 1 July 2019 / Accepted: 7 January 2020

Abstract
This article describes a heuristic argument for understanding certain physical systems in terms of properties that resemble the beliefs and goals of folk psychology. The argument rests on very simple assumptions. The core of the argument is that predictions about certain events can legitimately be based on assumptions about later events, resembling Aristotelian 'final causation'; however, more nuanced causal entities (resembling fallible beliefs) must be introduced into these types of explanation in order for them to remain consistent with a causally local Universe.

Keywords Folk psychology · Intentional stance · Teleology

1 Introduction

Neurotypical humans have a strong propensity to distinguish between animate and inanimate objects, and to understand humans and animals in terms of "folk psychology": beliefs, desires, moods, and so forth. Presumably, this proficiency provided an adaptive advantage for our evolutionary ancestors, by allowing better coordination of interactions (both cooperative and antagonistic) with other humans and animals.

From a common-sense perspective, the value of folk psychology is obvious: it is a very good (though imperfect) model of how other humans (and animals) actually behave. But this raises an interesting question: *why* are the specific constructs of folk psychology (belief, goal, perception, mood, etc.) a good way to understand other biological systems? Is our folk-psychological competence narrowly tailored for the specific complex systems that just happen to have dominated our evolutionary environment, or does it capture something more generally applicable? To put the same question another way, what are the basic abstract properties required for folk psychology to be a good model of a system?

✉ Ron Chrisley
ronc@sussex.ac.uk

Simon McGregor
s.mcgregor@sussex.ac.uk

[1] Centre for Cognitive Science, Sackler Centre for Consciousness Science, and Department of Informatics, University of Sussex, Falmer, United Kingdom

Published online: 09 March 2020

The philosophical discourse in this area typically pairs 'beliefs' with 'desires'; however, the form of our argument is simplified if we consider desires for specific attainable outcome events, rather than generic preferences which might apply to extended temporal states of affairs. We will therefore talk about 'goals' rather than 'desires'.

This article argues that while some folk-psychological notions may be ecologically narrow, the utility of a belief-goal logic arises naturally from some circumstances that are easy to characterise. Indeed, the formal structure of belief-goal explanations is the most rational way to understand certain sorts of system, even for a theorist who entirely lacks animistic or folk-psychological intuitions. Essentially, these systems are ones where the theorist's expectations regarding the system's longer-term behaviour provide useful traction on predictions of the system's shorter-term behaviour, and the theorist can additionally make use of information regarding causal constraints: specifically, knowledge about what environmental factors can affect the system's behaviour.

We begin in Sect. 1.1 with some brief comments on belief-goal explanations, and in Sect. 1.2 on causation. Section 2 introduces the notion of an 'ultra-scientist': a hypothetical theorist who lacks belief-goal intuitions, and illustrates this concept by reference to the famous Sally–Anne test in developmental psychology. Having introduced the ultra-scientist, we proceed to describe the methods and results of a hypothetical experiment in Sect. 3. Based on some simple physical constraints, Sect. 4 argues that the ultra-scientist can apply what we call a 'temporal-interpolation' mode of reasoning to her experimental results that we describe as a teleological stance. We use the term teleological because it does not make use of any notion resembling belief.

Section 5 introduces a revised version of the experiment, by adding contextual information that entails some reasonable inferences regarding the causal structure of events; these causal constraints permit the ultra-scientist to employ an augmented temporal-interpolation mode of reasoning, which involves considering what behaviour would produce a particular outcome in a potentially counterfactual world, even if this behaviour would produce a different outcome in the real world. We suggest that this mode of reasoning reproduces the core structure of belief-goal reasoning, with the counterfactual world playing the role of a belief, and the outcome in the counterfactual world playing the role of a goal.

Finally, Sect. 6 provides a brief discussion, Sect. 7 a comparison with related work, and Sect. 8 our conclusions.

1.1 Folk Psychology

Folk psychology is the term used by philosophers to refer to our implicit (nonscientific) theory of how other human beings (and, to some degree, non-human animals) operate. When we are asked to explain why a particular person behaved in a particular way, we respond by using the vocabulary of folk psychology: intent, free will, force of habit, emotion, and so forth. 'Folk science' theories such as folk psychology and folk physics are indispensable in the everyday life of the neurotypical human;

they are fallible, but provide enough predictive purchase to support human social competences.

This article will consider some very basic notions from folk psychology (goals and beliefs), and argue that their logic can be derived from rather general scientific explanatory principles. It is probably no coincidence that goals and beliefs are (more or less) the same concepts underlying Dennett's famous intentional stance originally detailed in Dennett (1987), which he succinctly defines in Dennett (2009):

> The intentional stance is the strategy of interpreting the behavior of an entity (person, animal, artifact, whatever) by treating it as if it were a rational agent who governed its 'choice' of 'action' by a 'consideration' of its 'beliefs' and 'desires.'

We will note that Dennett's definition can serve only as a skeletal structure; we do not attempt here to construct a definition of rational action, but we do appeal in Sect. 5.3 to some intuitive ideas about goals and beliefs.

The aim of this paper is to characterise the properties of situations that call for belief/desire—like explanations, in a 'non-intentionally epistemic' manner: that is to say, in terms of the non-intentional properties a system is understood to have by a theorist, and the pragmatic reasoning capacities of that theorist regarding those properties. We do not merely wish to defend folk psychology as ineliminable in complex social situations, like Fodor (1978) and Horgan and Woodward (1985); or like Dennett to differentiate between the physical and intentional stances; we wish rather to explore what aspects of the intentional stance are rationally mandated by the physical stance, and under what circumstances they are so mandated.

The particular folk-psychological principle that we intend to derive an analogue of is the following:

Belief-Desire Logic An agent X should be expected to behave in ways that would bring about its goals ω if its beliefs Y_1 were true.

We will get there in two stages. Firstly, we will justify the analogue of a simpler principle, which resembles the reasoning attributed to people who supposedly lack a theory of mind:

Desire-Only Logic An agent X should be expected to behave in ways that will bring about its goals ω in the actual world.

We will show that an analogue of this "teleological" mode of reasoning can be derived from simple temporal interpolation. Then, we will explain how the understanding of causal constraints can be used to refine this mode of reasoning, necessitating the introduction of a potentially counterfactual state of affairs Y_1 as an additional explanandum for behaviour, and resulting in more powerful predictions.

Note that if we were aiming to provide a completely non-intentional account of belief-desire logic, we would need to explain in virtue of what properties the theorist themselves merited an intentional description. We are not concerned here with

resolving this regress: while we will permit the hypothetical theorist only to deploy the tools of the physical stance, we will allow ourselves to characterise the theorist intentionally.

1.2 Causation

To make our case, we will find it useful to talk of sensors and actuators. Our definitions of sensors, actuators and rationality invoke the notion of causation. How to naturalise causation is a contended topic; we favour intervention-based statistical models of causality (e.g. Pearl 2000), which consider both the properties of the system Ω in the absence of external interference, and also Ω's properties when disturbed by specific external interventions that are not correlated with any of Ω's variables.

The intervention-based notion of causation allows us to focus on what would happen if the agent's 'externally visible' (actuator) variables were set to arbitrary values by an external intervention. Hence, we do not consider any constraints posed by the agent's unknown and/or incomprehensible internal mechanisms. Such constraints are what presumably inform subtler notions in folk psychology, such as forgetting; we are not primarily concerned with them in the present article, but rather with more fundamental concepts resembling beliefs and goals.

2 The Ultra-Scientist

As an aid to imagination, this article proposes a notion of an "ultra-scientist": a hypothetical empirical scientist who does not have folk-psychological intuitions. The ultra-scientist entirely lacks the ability to apply the vocabulary of agency and mind to other systems, and can only reason about their physical behaviour. In Dennettian terms, we assume that she can apply the physical stance but is unfamiliar with the intentional stance. This rhetorical device is intended to sharpen our argument that explanatory posits resembling beliefs and desires are rationally mandated by certain 'non-intentionally epistemic' relations between a theorist and an explanandum.

Note that this article will take some logical liberties in adverting to the notion of an ultra-scientist: for instance, the postulate of agency-blindness probably precludes a capacity for appropriate linguistic behaviour, but we will still consider 'what the ultra-scientist would say'. Along these lines, one possible objection to our argument is that the practice of science intrinsically requires a mastery of the intentional stance (at least to the level of competence with belief-desire explanations), and that therefore the notion of an ultra-scientist is fundamentally inconsistent. We will discuss this possible objection in Sect. 6.1.

2.1 Physical and Intentional Concepts

Put roughly, our thesis is that (core aspects of) intentional reasoning can be derived directly from ordinary physical reasoning, without any need for additional precepts. But this rough formulation invites misunderstanding. For example, put that way, our

thesis might invite the thought that we are supposing our ultra-scientist to possess no intentional (or proto-intentional) concepts (that is, no concepts *of* intentionality or proto-intentionality; in a sense, all concepts are "intentional concepts" in that concepts exhibit intentionality, but that is not the sense we are using here). But it would be a mistake for us to agree with that characterisation of our thought experiment, because we wish to leave open this possibility: that in virtue of possessing the ability to (unwittingly) match the successes of proto-intentional reasoning, the ultra-scientist might thereby be said to possess concepts of proto-intentionality, at least implicitly.

A distinction may be of use here, between concepts of things that are obviously naturalisable (i.e., things for which no naturalisation question arises), and concepts of things the naturalisability of which is not obvious (i.e., things for which it is reasonable to ask: is this thing naturalisable?). The only concepts that we are stipulating to be possessed by the ultra-scientist are of the former sort: ones that we all readily agree are of things that are non-problematically naturalistic. If our argument is successful, it will also be the case that the ultra-scientist, by virtue of possessing these concepts, will thereby possess concepts of the latter kind: concepts of proto-intentionality, for which the question of naturalisation arises (at least before our argument and analysis!). But by taking care, we avoid begging the question; we avoid *assuming* that the ultra-scientist possesses no concepts of proto-intentionality, and we likewise avoid *assuming* that the ultra-scientist already possesses concepts of proto-intentionality. Thus we leave open the possibility of a substantive, naturalistic explanation of (concepts of) proto-intentionality.

2.2 The Sally–Anne Test

We will illustrate the ultra-scientist's imagined deficits, by reference to the famous Sally–Anne test as an example. In this experiment, a scene is acted out with two dolls called Sally and Anne. Sally places a marble in a box, and then leaves the scene. While Sally is absent, Anne moves the marble to a different box. When Sally returns, the participant is asked "Where will Sally look for her marble?" This test is difficult for young children, who tend to point to the current location of the marble instead of the original location; Sally ought to believe that the marble is in the original location.[1]

The ultra-scientist 'doesn't even fail' the Sally–Anne test. The standard wording of the question ('Where will Sally look for her marble?') is incomprehensible to her, since she understands only naturalistic physical concepts, and these do not include a notion of 'looking for'. The question must be reworded as "Which box will the Sally object's immediate trajectory reduce its distance from?", to which she will sensibly answer "I do not know the internal mechanisms of the Sally object, and have not made enough observations to predict its behaviour." This same response will occur

[1] There is dispute over whether this difficulty actually reflects an inability to imagine false beliefs, or an artefact of a linguistic prompt. See, e.g. Scott and Roby (2015) for a discussion.

whether the scene is acted out with dolls or with actual humans; for the ultra-scientist, both are merely objects constituted of ordinary physical matter.

We will argue that, despite having no folk-psychological intuitions, considering some relatively simple empirical properties of our physical environment would lead the ultra-scientist to develop ideas similar to the folk-psychological concepts of belief and goal.

3 The Ball and Blob Experiment

This article will consider an imaginary behavioural experiment performed by an (ultra-)scientist. Note that this experiment is not intended to resemble anything in the existing scientific literature; it is a thought experiment intended to illustrate where reasoning objectively about certain sorts of behaviour from physical first principles would lead.

The scientist has been given a peculiar blob, measuring 1 cm across, which she has never encountered before. The scientist is instructed to perform an experiment as described below.

3.1 Setup

The experiment involves a ball-bearing rolling down a rugged surface. The ball can be placed in a housing mechanism at the top of the surface; this mechanism can be triggered with variable parameters to release the ball at a specified angle and velocity. The surface then channels the ball into one of two chutes which release the ball vertically above a solid floor. The ball takes no more than 5 s to exit a chute, and the chutes are 10 cm apart (see Fig. 1).

The surface is complicated, so that the final position of the ball is sensitively dependent on its initial release parameters, but a certain set of parameters has been discovered that produce reliable end outcomes even when the mechanism closing the box door is triggered.

The floor is divided into two halves by a fine visible line, perpendicular to the axis between the chutes. A laser mechanism allows the position and velocity of objects on the floor to be tracked.

3.2 Protocol

The experiment consists of a number of trials, performed by an automated mechanism. In each trial, the system is initialised by returning the ball to its housing, and positioning the blob at a fixed location on the floor in front of the slope. These events occur automatically, without manual intervention. The following events then occur:

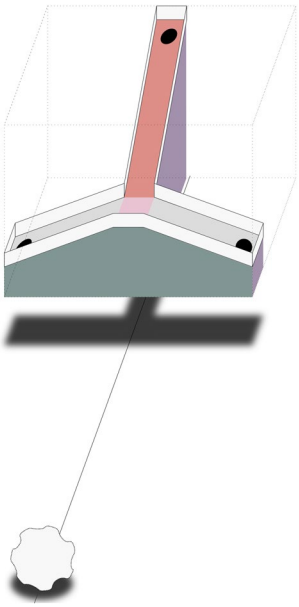


Fig. 1 The ball and blob experiment setup. Dotted lines indicate perspective

1. When the blob has been repositioned, the housing is triggered with one of the known-outcome parameters, causing the ball to be released almost instantaneously.
2. 6 s later, the experiment ends.
3. Several variables are recorded during the experiment, and become available to the scientist as data:

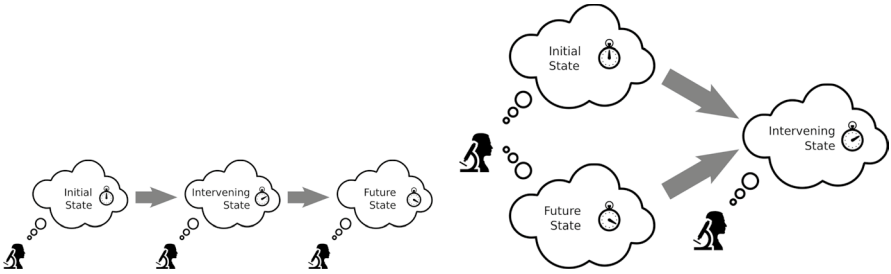
    a. The identity of the chute through which the ball fell;
    b. The time at which the ball fell through the chute;
    c. The maximum speed of the blob during the entire trial;
    d. The final position of the blob.

Note that the scientist has access only to the data recorded in this manner.

3.3 Results

The scientist performs the experiment, and obtains the following results:

1. The ball never takes more than 5 s to fall through a chute.
2. In over 90% of trials, within 5 s the blob ends up directly underneath the chute from which the ball exited.
3. In no trial does the blob's maximum speed ever exceed a specific speed, which we will call $v_{\max}$ cm/s.



(a) Forward inference **(b)** Temporal interpolation inference

Fig. 2 Inference from the present forwards mechanically (**a**) and using temporal interpolation (**b**). Forward inference resembles Aristotelian 'efficent explanation' or Dennett's 'physical stance'. Temporal interpolation resembles Aristotelian 'final explanation' or Dennett's 'design stance' and 'intentional stance'

4 'Temporal Interpolation' vs. 'Mechanism-Forward' Reasoning

Remember that the scientist has no direct data regarding the blob's overall trajectory, and knows only its fixed initial position and its recorded final position in each trial. She begins a new trial, and now wishes to predict which side of the line the blob will be on after time 0 plus 3 s.

In this section, we will demonstrate how the scientist can use the data from the experiment to predict some aspects of the blob's behaviour at time $t = 3$ by making use of her knowledge about:

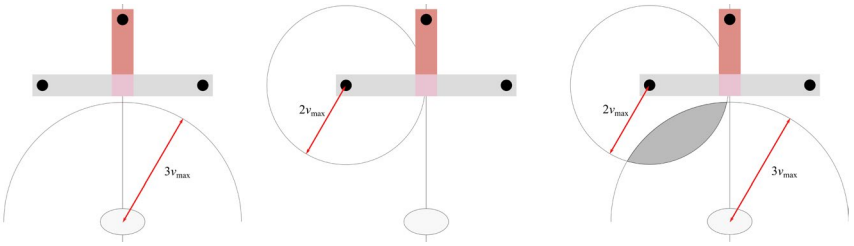
1. Aspects of the blob's initial state at time $t = 0$;
2. Aspects of the blob's expected state at time $t = 6$;
3. Some constraints on how the blob can behave between these times.

We call this mode of reasoning 'temporal interpolation' simply to emphasise that it depends on assumptions about the later effects (in this case, the effects at $t = 6$) of the behaviour being predicted (in this case, the state at time $t = 3$) *as well as* information about the system's prior state (in this case, the state at time $t = 0$).

We propose that this temporal-interpolation mode of reasoning constitutes a sort of teleological stance, and that the contrast between temporal-interpolation and mechanism-forward explanations somewhat resembles the contrast between Aristotle's 'final' and 'efficient' explanations.

4.1 Aristotelian Explanations

The ancient Greek philosopher Aristotle famously distinguished between four different kinds of explanation ('four causes', in traditional translations). His last category is that of 'final causes' or 'final explanations': roughly speaking, explaining something in terms of the ends that are produced by it. In this sense, we suggest that temporal-interpolation reasoning resembles final causation, while mechanism-forward reasoning resembles efficient causation (see Fig. 2).



(a) Start Position **(b)** End Position **(c)** Combination

Fig. 3 Deducing intermediate states from initial and final states. **a** The positions the blob can reach in 3 s from its starting position. **b** The positions from which it can reach its final position in 2 s. The shaded area in **c** shows the area that the blob must be 3 s from now, if it is now in the depicted position, and ends up under the left-hand chute 5 s from now, with its speed never exceeding v_{max}

There are, of course, various views on how final causation might be interpreted and applied as a principle within cognitive science. For instance, Rachlin's *teleological behaviourism*, like our approach, emphasises the future context of behaviour:

> One important feature of the behavioral viewpoint bears emphasis: [t]he question [of the motivation behind behaviour] may be settled by reference to future as well as to past events, because the temporal context of a brief event extends into the future as well as the past.

However, in Rachlin (1992), he considers final causation to be a constitution relation: "[A] final cause is to its effects as a wider concept (like a dance) is to the particulars (like the steps) that make it up." Nothing in our argument rests on its relation to Aristotelian thought, so we do not insist on any particular interpretation of final explanation in the Aristotelian sense.

4.2 Temporal-Interpolation Reasoning

From her empirical predictions of the further future, and an assumption that the blob remains a single solid lump of matter, the scientist can make certain inferences regarding the near future. Using induction, she extrapolates from her past empirical data in the ordinary scientific fashion. Hence she predicts the following:

1. The blob will end up under the chute from which the ball exits within 5 s.
2. The ball will take no more than 5 s to exit the chute.
3. The blob's maximum speed will be v_{max} cm/s or less.

Figure 3 shows the possible locations that the blob might be in after 3 s, according to these constraints. The scientist concludes that

4. After 3 s, the blob will be on the same side of the line as the chute that the ball will eventually exit from.

This reasoning resembles a *boundary value problem* in the mathematical sciences, i.e. a problem in which the shape of a curve is deduced from knowing the beginning and end points of the curve (along with some additional constraints on the curve, usually in the form of differential equations).

4.3 Constrasts with Mechanism-Forward Explanations

It is illuminating to contrast this temporal-interpolation reasoning with the mechanical type of explanation more common for physical systems: if the scientist knew more about the blob's internal mechanisms, she could perhaps have reached the same conclusion (the blob will be on the same side of the line as the ball) by reasoning forward from its initial state and its environment's initial state. For instance, suppose that the blob followed a predictable trajectory such that its position x_{n+1} at time $t = n + 1$ was some function $f(x_n, \lambda)$ of its position x_n at time $t = n$ (as well as the ball release parameters λ). The scientist would only need to know x_0 and λ, and could calculate the blob's position x_3 at time $t = 3$ via a series of intermediate computations: $x_1 = f(x_0, \lambda)$; $x_2 = f(x_1, \lambda)$ and finally $x_3 = f(x_2, \lambda)$.

The temporal-interpolation explanation and the mechanism-forwards explanation are equally scientifically valid, but they make use of different sorts of knowledge: in the mechanism-forwards case, a detailed knowledge of covering laws over a shorter time-frame than the event being explained; in the temporal-interpolation case, a knowledge of covering laws over a longer-time frame, plus knowledge of some relevant constraints over the interval ending in the event. In the mathematical case, the equations that need to be solved are even of a different sort: the mechanism-forwards case usually involves solving a definite integral, while the temporal-interpolation case usually involves solving a constrained differential equation.

Note that the constraints from the scientist's data may not determine on which side of the line the blob will be found during the first few seconds of each trial; they may leave leeway for the blob to start moving on the other side of the line, or switch back and forth. This would usually not be true for a mechanism-forward argument.

4.4 Simplicity of Temporal-Interpolation Reasoning

In principle, the scientist does not have to describe her original data in terms of where the ball ends up. After all, we have stipulated that the ball's final position is reliably predictable from the initial angle and velocity with which it is released. Hence, the scientist could just as well say, "when the ball is released with parameters so-and-so, the blob ends up under the left chute; otherwise, under the right chute".

We make this stipulation in order to demonstrate that the scientist can have reason to use a temporal-interpolation mode of reasoning, even if in principle the mechanism-forward mode of reasoning is also available. Recall that the surface is rugged and complicated, and that the ball's trajectory is sensitively dependent on initial conditions. In consequence, the set of initial conditions under which the blob ends up under the left chute will effectively be an arbitrary list of numbers with

no intuitively distinguishable pattern. The description "the blob ends up under the chute that the ball falls through" is a much more elegant and succinct one.

Note that if the relation between the ball's initial state and the blob's final state were a transparent one, there would be no need for a temporal-interpolation (teleological) explanation. For instance, suppose that the ball's final position depended simply on whether its initial velocity took it to the left or the right. Then it would be simple to describe the blob's behaviour using forwards (mechanical) reasoning: the blob will end up under the left chute if the ball is released to the left, and under the right chute if the ball is released to the right. In other words, it is because the final conditions are easier to describe than the initial conditions that a teleological description becomes convenient.

4.5 The Teleological Stance

The 'final explanation' character of temporal-interpolation reasoning contrasts with the 'efficient explanation' of mechanism-forward reasoning in a way that is reminiscent of Dennett's distinction between the design (or intentional) stances and the physical stance. However, we have so far presented no convincing reason for an ordinary observer, let alone an ultra-scientist, to construct an intentional-like explanation for describing the blob's behaviour; perhaps the blob is simply drawn by some magnetism-like effect to the ball.

Indeed, the availability of long-term predictions with short-term constraints, and hence of the consequence-backward mode of reasoning, suffices only to justify what we will call a *teleological* stance: one that involves goal-like but not belief-like constructs.

A teleologist could be posed the Sally–Anne question, and would "fail" the test, as follows. The long-term prediction is that, when Sally is seeking her marble, she will end up with it in her hand soon. Using reasoning resembling that shown in Fig. 2b and described in Sect. 4.2, the teleological conclusion is that Sally's trajectory will take her towards the box with the marble in it (and, since neither her hand nor the marble can pass through solid matter, that the box will soon be open).

In the next section, we will see that making additional assumptions, involving how chains of causation work in a system, can lead to a more sophisticated mode of temporal-interpolation reasoning, that introduces objects resembling false beliefs.

5 Causation and Counterfactual Final Explanations

The scientist has based her blob/environment distinction on purely phenomenal considerations: the blob looks like a coherent lump of stuff, albeit one that moves in peculiar and spontaneous ways. Consequently, she has some notion of where, at time t, the blob X ends and the environment Y begins. This corresponds to a closed spatial boundary.

Because the scientist believes our Universe to be spatially local (i.e. that all long-range interactions are mediated by shorter-range ones), she will assume that all

causal interactions between the blob and its environment must be mediated by variables on either side of this boundary.[2] This knowledge can be used to infer additional information about the blob's likely behaviour in different circumstances.

5.1 Mediating Causal Variables

Let's define a set of variables around the blob's spatial boundary, which we will call S. These variables form a subset of the entire environment state Y. We will write $S_{[t_1 \cdots t_2]}$ to denote the trajectory of these variables over the interval between t_1 and t_2.

We will suppose that the variables S causally mediate the causal effects on X produced by other variables $\bar{S}$ in the environment. This can be given a technical scientific meaning in terms of causal Bayesian networks (Pearl 2000). These networks are based on structural causal models, which describe the consequences of hypothetical external interventions in a system: for instance, sprinkler activity may be statistically correlated both with hot weather and with a wet pavement (sidewalk), but an intervention that turns the sprinkler on when the weather is dull will only make the sidewalk wet; it will not change the weather. Such relationships are already captured 'for free' in dynamical systems formulations such as ordinary differential equations (Fig. 4).

5.2 Reasoning from Causal Constraints

Suppose the scientist conducts a new experiment: she starts the trial as usual. The ball is released with some angle and velocity that, in normal conditions, reliably result in its exiting the box from chute 1. However, at time t_1, she applies a magnetic field inside the box, thus making conditions abnormal, and causing the ball to eventually exit from chute 2. Which side of the line should she expect the blob to be on at time t_2 (time 0 plus 3 s)?

On the information we have presented so far, she should follow the reasoning presented in Sect. 4, and conclude that the blob should be found on the chute 2 side of the line at time t_2, since the blob tends to end up where the ball does, and the ball will end up in chute 2. However, the scientist believes that

1. It is plausible that applying the magnetic field between t_1 and t_2 has no relevant causal effect on $S_{[t_1 \cdots t_2]}$ (the boundary around the blob between closing the box and the 3-s mark) or on the blob's internal state X_{t_1} at time t_1.

She bases this belief on some additional knowledge that we have so far omitted to share with the reader (see Fig. 4):

[2] Note that even effects such as gravity are assumed to be 'carried' by particles that propagate through space at a limited velocity; at a small enough spatiotempioral scale, such effects are still mediated by the boundary states.

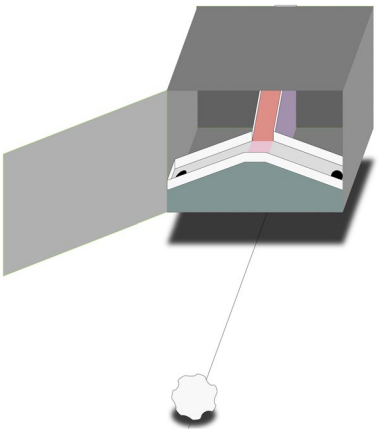


Fig. 4 The true ball and blob experiment setup

2. The sloped surface is inside a box, whose front side is initially open, but is closed by a remote mechanism, becoming completely closed shortly before time t_1.
3. The box is lead-lined, and this blocks most electromagnetic causal influences.
4. The entire experiment is conducted in an evacuated vacuum chamber, and the box is suspended by a few sturdy cables. This makes vibration-based causal influences unlikely.
5. The vacuum and materials make diffusion-based causal influences unlikely.

If she is right about the magnetic field having no causal influence on the blob, then a wholly new argument must obtain:

6. By definition, the combination of the blob's initial state X_{t_1} and its 'sensory' trajectory $S_{[t_1 \cdots t_2]}$ suffice to specify the blob's motion trajectory[3] over the interval $[t_1 \cdots t_2]$. This is because its 'sensory' trajectory was defined to consist of everything external that could causally affect the blob.
7. Therefore, if applying the magnetic field at time t_1 does not have a relevant effect on $S_{[t_1 \cdots t_2]}$ or X_{t_1}, it cannot have an effect on the blob's motion trajectory during that time.
8. Therefore, the blob should be expected to behave in the same way as if no magnetic field had been applied, and (following the reasoning presented in Sect. 4) should be expected to be on the chute 1 side of the line at time t_2.

Of course, the scientist does not know for sure that there will be no causal effect on the blob when the magnetic field is applied. Intuitively speaking, the blob might still be able to track the ball after the lead box is closed, e.g. through the gravitational effects of the ball or through some currently unknown physical principle.

[3] Or, if the blob's behaviour is not deterministic, a probability function over possible motion trajectories.

We will suppose, however, that the ultra-scientist's reasonable guess is supported by the behavioural evidence, in that the blob is indeed insensitive to the magnetic-field intervention, once the box has been closed. Note that if we were looking for reasons to treat the blob as a system capable of false beliefs, this would be precisely the sort of evidence we would want to see.

The introduction of causal-equivalence constraints into the scientist's reasoning represents a qualitative advance on the simple temporal-interpolation 'boundary value' explanations described in Sect. 4. It allows the scientist to 'go against the statistics', correctly predicting atypical outcomes (the blob will end up at a different chute than the ball) based, nevertheless, on those very same statistics (the blob typically ends up at the same chute as the ball).

5.3 Proto-Intentionality

In the previous section we discussed atypical circumstances (the application of a magnetic field) for which certain future outcomes (being under the ball when it falls) can be changed, while plausibly having no causal effect on the blob's behaviour in the intermediate future (which side of the line it is on after 3 s). We showed how the scientist can combine causal-equivalence principles with temporal-interpolation reasoning to correctly predict that the blob will approach the 'wrong' chute in such circumstances.

There is an interesting physical principle here: suppose a 'typical' environment of type Y_1 and an 'atypical' one of type Y_2 are causally indistinguishable in their effects on an object X during some time interval $T = [t_1 \cdots t_2]$, and an event ω typically occurs at time t_2. Then, in Y_2-type environments, X is likely to behave during T in ways that are compatible with ω occurring in Y_1-type environments, even if that behaviour prevents ω from occurring in its actual Y_2-type environment.

If we identify the object X with an agent, the external causal influences S on X with sensory influences, the environment class Y_1 with a belief, and the event ω with a goal, we essentially reproduce the following principle from folk psychology:

1. An agent X should be expected to behave in ways that would bring about its goals ω if its beliefs Y_1 were true,

where X's beliefs Y_1 equate to the most 'typical' state of affairs compatible with its sensory influences to date. It is interesting to compare this principle with Dennett's characterisation of beliefs in the intentional stance:

> A system's beliefs are those it *ought to have*, given its perceptual capacities, its epistemic needs, and its biography... "ought to have" means "would have if it were *ideally* ensconced in its environmental niche."

(Dennett 1987, p. 49, italics in original). We have effectively replaced Dennett's notion of an 'environmental niche' with the weaker notion of a typical environment in which some future event is readily predictable.

6 Discussion

We have argued, in effect, that the following causal conditions are sufficient to mandate a highly limited sort of teleological explanation.

1. Given some set of conditions c at time t_{initial}, a system Y behaves in a way that tends to cause some instantaneous outcome state g at a later time t_{final}.
2. At time t_{predict} (with $t_{\text{initial}} \leq t_{\text{predict}} \leq t_{\text{final}}$), the most practical way to predict Y's behaviour at time t_{predict} is to infer it from the constraints imposed by c and g together.

Condition 2 constitutes the justification for a teleological or 'final' explanation (*a la* Dennett's 'design' and 'intentional' stances), rather than a mechanistic or 'efficient' explanation (*a la* Dennett's 'physical' stance).

By stipulating an additional condition, we introduce a mandate for entities resembling false beliefs:

3. Suppose an event E can occur at an intervening time t_{event}, such that:

    a. There are no causal mechanisms by which E can affect Y's behaviour until time t_{final}; and
    b. With Y's behaviour held constant, E will cause g not to occur.

In such circumstances, we can consider a counterfactual Universe U' in which E did not occur, and we should expect Y's behaviour from time t_{event} to time t_{final} to be that behaviour which would cause g to occur if the system were in Universe U'. In other words, to behave as if it believed it was in U' and wanted to achieve g.

This argument is intended to show that belief-goal explanations are mandated by logically general features of our environment, not narrow *ad hoc* ones. We have argued that the above criteria are sufficient conditions, not that they are sufficient and necessary; perhaps there are even simpler empirical conditions that would call for an intentional explanation.

6.1 The Ultra-Scientist

We have made use, in this article, of a fictional creature who is forced to 'reconstruct' certain aspects of the intentional stance from purely physical stance principles, by virtue of her natural unfamiliarity with the intentional stance (at least, as regards systems external to herself). In Sect. 2, we mentioned the possibility that this rhetorical device may be fundamentally inconsistent. However, the question at issue is whether the bare bones of belief-desire reasoning (as we have described it here) can be derived from purely physical stance principles. Hence, our argument's validity rests on two questions:

1. Whether the reasoning we impute to the ultra-scientist (regarding the blob's behaviour) in Sects. 4.2 and 5.2 constitutes sound reasoning; and
2. Whether that reasoning requires intentional-stance premises to be assumed in addition to physical-stance premises.

We contend that the answer to question 1. is "yes" and question 2. is "no".

6.2 Limitations

Our argument is a simple indicative one with limited scope, and its form is heuristic, rather than mathematically rigorous. In Appendix A, we provide a short proof supporting the integrity of the 'temporal-interpolation' argument presented in Sect. 4.

The argument is meant only to help elucidate the circumstances under which differing forms of predictive reasoning are useful. We have deliberately refrained from commenting on what, if anything, the use of a temporal-interpolation, causally-constrained mode of reasoning implies about beliefs or goals in the system itself.

While the properties that we have identified as supporting belief-goal explanations are rather generic ones, we have not thereby explained why systems with these properties exist in our environment. That issue touches on questions about the origins of life, and about the structure of human cognition, which are well beyond the scope of the current article.

6.3 The Clever Blob

We may parenthetically consider what conclusions we would draw about the blob if we, armed with the full power of psychological insight, were in the ultra-scientist's position. If the scientist is correct, and the blob's behaviour is causally unaffected by the ball's trajectory after t_1 (i.e. it ceases to have any information about the ball's trajectory after this point), the blob must be capable of somehow translating information about the ball's initial trajectory into information about the chute from which the ball will exit, despite the complex surface that the ball rolls over.

This is in curious counterpart to the scientist's stipulated cognitive limitations: remember that, as discussed in Sect. 4.4, if the theorist can transparently describe the set of release parameters which send the ball down the left-hand chute, there is no need for her to pick out the future consequences as a point of reference for describing the blob's behaviour. Exploring the relationship between a theorist's own cognitive limitations and the capacities of systems they understand as cognitive is an interesting topic we leave for future research.

7 Comparison with Related Approaches

7.1 Life, Mind and Free Energy

Our treatment allows a novel perspective on autopoietic notions of life and mind (Thompson 2007), including autopoietic aspects of the so-called 'free-energy' framework (Friston 2013). These seek to ground notions of purpose and/or cognition in the self-maintenance of certain sorts of physical system, known as autopoietic systems (Maturana and Varela 1987). The free-energy framework goes further by considering a specific statistical mechanical property known as an 'ergodic Markov blanket'.

Recall condition (1) from Sect. 6:

1. Given some set of conditions c at time t_{initial}, a system Y behaves in a way that tends to cause some instantaneous outcome state g at a later time t_{final}.

Autopoietic systems, and systems separated from their environment by an ergodic Markov blanket, are special cases for which this condition is met. For autopoietic systems, the system's existence at time t_{initial} tends to cause its existence at time t_{final}. Ergodicity imposes an even stronger constraint: not only will the system still exist at time t_{final}, but the 'inputs' from its environment can be predicted from a specific (ergodic) distribution.

If the dynamics of the systems in question are sufficiently opaque, meeting condition (2) from Sect. 6, we thereby have an argument from first principles that autopoietic systems must appear to seek their own survival, and ergodic systems must apparently seek to minimise their free energy.

7.2 Dennett and the Intentional Stance

Dennett introduces the (non-folk-psychology-using) Martian super scientist as a way of expressing an *objection* to the intentional stance (that is, to his intentional systems theory based on the intentional stance)—the objection that the intentional stance is observer-relative. The objection goes something like this: in order to deal with an earlier objection (the lectern objection), Dennett stipulates that a belief-desire possessing system is one for which the intentional stance is required in order to reliably and voluminously predict its behaviour. On such a view, we come out as "true believers", while a lectern does not: even though the intentional stance can be used to predict a lectern's behaviour, the physical stance does just as well, if not better. But the Martian super-scientist is stipulated to be such that it would not need the intentional stance to predict a typical person P's behaviour, even if we need the intentional stance to do the same. So is P a true believer or not? It seems that Dennett's theory would make whether or not a subject is a true believer to be a subjective matter: P is a true believer relative to us, but relative to the Martian super-scientist, not.

Dennett's counter to this objection is that there are patterns that the super scientist would not be able to explain without folk psychology:

"The Earthling and the Martian observe (and observe each other observing) a particular bit of local physical transaction. From the Earthling's point of view, this is what is observed. The telephone rings in Mrs. Gardner's kitchen . She answers, and this is what she says: "Oh, hello dear. You're coming home early? Within the hour? And bringing the boss to dinner? Pick up a bottle of wine on the way home then, and drive carefully." On the basis of this observation, our Earthling predicts that a large metallic vehicle with rubber tires will come to a stop on the drive within one hour, disgorging two human beings, one of whom will be holding a paper bag containing a bottle containing an alcoholic fluid . The prediction is a bit risky, perhaps, but a good bet on all counts. The Martian makes the same prediction, but has to avail himself of much more information about an extraordinary number of interactions of which, so far as he can tell, the Earthling is entirely ignorant . For instance, the deceleration of the vehicle at intersection A, five miles from the house, without which there would have been a collision with another vehicle - whose collision course had been laboriously calculated over some hundreds of meters by the Martian. The Earthling's performance would look like magic! How did the Earthling know that the human being who got out of the car and got the bottle in the shop would get back in? The coming true of the Earthling's prediction, after all the vagaries, intersections, and branches in the paths charted by the Martian, would seem to anyone bereft of the intentional strategy as marvelous and inexplicable as the fatalistic inevitability of the appointment in Samarra".

A full understanding of what is going on in Dennett's counter is beyond the scope of this paper, but we can note here that our use of the ultra-scientist does have some similarities to Dennett's Martian. The main question of our paper is whether the ultra-scientist can make sense of the blob's ability to predict the behaviour of the ball, which is structurally similar to the question of whether Dennett's Martian can predict/explain/understand Mrs. Gardner's ability to predict the behaviour of the person on the phone (presumably Mr. Gardner). One idea of how these relate to each other is shown in Table 1.

But there is an important disanalogy: Mr. Gardner is a person, with beliefs and desires, whereas the ball in our example is not. The ball is playing the role of the kinds of things that intentional agents reason about, not that of an intelligent agent itself. Dennett's example conflates these two. Presumably, his example would work just as well if Mrs. Gardner were reasoning about inanimate objects (like balls),

Table 1 One understanding of how our thought experiment relates to Dennett's

Dennett	McGregor and Chrisley
Martian super-scientist	Ultra-scientist using only forward reasoning
Earthling	Ultra-scientist using temporal-interpolation reasoning
Mrs. Gardner	Blob
Mr. Gardner	Ball

Table 2 Another understanding of how our thought experiment relates to Dennett's

Dennett	McGregor and Chrisley
Martian super-scientist	Us
Earthling	Ultra-scientist using temporal-interpolation reasoning
Mrs. Gardner	Blob
Mr. Gardner	Ball

since it is the Earthling's (and opposed to Mrs Gardner's) successful ascription of beliefs and desires to others (as opposed to possessing beliefs and desires) that is supposed to puzzle the Martian. Our example eliminates this confusing conflation.

But there's another way of seeing how the two examples relate. Dennett argues that the Martian super-scientist would not be able to make sense of the Earthling's ability to predict Mrs Gardner's behaviour without seeing the Earthling as possessing the ability to ascribe the concepts of folk psychology to others. Similarly, we argue that we cannot make sense of the ultra-scientist's ability to predict the blob's behaviour without seeing the ultra-scientist as possessing the ability to ascribe proto-intentional relations to others. This interpretation is shown in Table 2.

Seeing things this way reveals an important difference of explanatory strategy between Dennett's example and ours. In stressing the conceptual distinctness of the intentional stance as an explanatory enterprise, Dennett at best failed to show how the things posited by the intentional stance are metaphysically continuous with those postulated by the physical stance. At worst, he introduced a dualism from which there is no recovery. Our goal, on the other hand, is to emphasise the continuity between intentional and non-intentional explanations. Thus, our strategy (as clarified in Sect. 2.1) was to present the blob, the blob' s behaviour, and the ultrascientist's mode of reasoning about the blob, in physically-grounded terms, without explicitly stipulating that any of them are intentional, so that such a designation might be a conclusion of our analysis, rather than a presupposition of it.

It should be noted that it seems that Dennett only allows his Martian to use mechanism-forward reasoning (like the stereotypical version of Laplace's demon). This weakens his point considerably, because that somewhat arbitrary restriction has other consequences that undermine Dennett's points concerning intentionality. For example, for a Martian that is limited to mechanism-forward reasoning, we can construct a scenario in which it is equally puzzled by the Earthling's ability to predict the behaviour of a rock rolling down a rugged slope. The Earthling takes one look at the slope, sees that there is nothing on which the rock can get stuck, and predicts that the rock will cross a line painted, half-way down, horizontally across the slope, knowing that the rock will end up at the bottom. The Martian simulates the rock's precise trajectory from a detailed knowledge of its shape and the slope's surface, and concludes that the Earthling is miraculously correct, despite the Earthling being ignorant of the details that permit a mechanism-forwards prediction. To avoid this situation, we need to grant the Martian the ability to engage in temporal-interpolation reasoning as well. Fortunately, doing so does not force us to attribute to the Martian naturalistically problematic concepts (see Sect. 1.2). Further, even with

enough temporal-interpolation reasoning abilities to find the Earthling's reasoning about the rock, slope, and the painted line non-miraculous, the Martian does not thereby necessarily have the kind of temporal-interpolation reasoning that is intentionality-invoking, as does our ultra-scientist.

Put another way, Dennett individuates explanatory practices by the concepts they use, which is fine if such concepts are already well-established, but dubious if there is no such prior grounding, and circular if one attempts to clarify those concepts by reference to the very explanatory practices which they are meant to help individuate. Such circularity is on display in Dennett's treatment of the intentional stance; the stance itself is defined in terms of belief and desire:

"Here is how it works: first you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in most instances yield a decision about what the agent ought to do; that is what you predict the agent will do."

But then the notion of belief is clarified in terms of the intentional stance:

"What it is to be a true believer is to be an intentional system, a system whose behaviour is reliably and voluminously predictable via the intentional strategy."

By contrast, in this paper we individuate explanatory practices in the first instance by the kinds of physical systems and physical behaviours they can and cannot predict/explain - this is why we take such care to tell you what the blob does in non-intentional terms, rather than saying something like "Suppose the ultra-scientist is trying to explain the behaviour of an intentional agent". This allows us to stipulate an empirical discontinuity within the class of physically characterised systems, rather than to stipulate a conceptual discontinuity between physical and intentional modes of explanation.

There are several advantages to taking our approach. Its non-circularity allows it to at least potentially provide a true naturalisation of intentionality. It also would permit (future) investigation of what it would take for an understanding system/understood system dyad to transit from being a dyad that does not, to a dyad that does, require the understander system to use teleological concepts to predict the understood system's behaviour. That is, we could finally begin to offer a non question-begging answer to the question "what is it about a system that makes the intentional stance useful in predicting its behaviour?".

Further, by weakening the intentional concepts involved from beliefs and desires to belief-like and desire-like states, our approach is better suited to exploring the points at which intentionality and intentional accounts first start to get a grip. Such exploration can be done in a way that is data-led, rather than motivated by top-down, a priori constraints from the special case of full-fledged beliefs and desires. Accordingly, we are not forced to use complex, difficult to naturalise semantic structures, instead relying only on the notion of non-actual situations.

8 Conclusion

A prototypical mechanistic explanation relies on identifying regularities that form constraints on a system's instantaneous behaviour. These constraints are strong enough that the system's future trajectory can be predicted from what is known about its current state.

However, there are other strategies for predicting a system's behaviour, which make use of different regularities. We have argued that a "teleological", "temporal-interpolation" explanation of a system's behaviour is appropriate whenever regularities in the system's instantaneous relationship to its environment are easier to identify at a longer time scale than at the target time scale. Based on knowledge of the present and expectations of the distal future, events in between can then be 'filled in' using some constraints on instantaneous behaviour; in contrast with the mechanistic account, these constraints do not have to be strong enough to predict the proximal future from the present alone.

When a teleological account is augmented with further knowledge regarding causal relations, we have argued that behaviour in atypical circumstances can be correctly predicted by considering the behaviour that would result in typical consequences in a (counterfactual) typical environment. We have proposed that this principle strongly resembles belief-desire logic, with the expected future consequences playing the role of a goal, and the counterfactual typical environment playing the role of a belief.

Temporal-Interpolation Probability Reasoning

This appendix describes a simple mathematical principle for temporal-interpolation reasoning using probability theory, which differs formally from mechanism-forward Bayesian inference.

Recall that the ball is released with parameters such as angle and velocity that are reliably related to the chute that the ball eventually exits from. We will classify each particular set of release parameters into an 'equivalence class' according to the chute from which we expect the ball to exit, based on those parameters.

Suppose we define the following random variables: A, the equivalence class of ball release parameters; B, the chute the ball comes out through; C, the chute the blob ends up under; D, whether the blob's trajectory takes it left or right.

We'll assume: $\mathbb{P}(a_i) = 0.5$; $\mathbb{P}(b_i \mid a_j) = 0.99$ when $i = j$ (and hence 0.01 otherwise); $\mathbb{P}(B = C) = 0.9$; and $\mathbb{P}(c_i \mid d_j) = 0$ when $i \neq j$, with $i,j \in \{0,1\}$, where 0 represents left and 1 represents right. These assumptions essentially state that

1. The ball is equally likely to be released with parameters that reliably make it 'fall left' as with parameters that make it 'fall right'.
2. When the ball is released with 'fall left' parameters, it is 99% likely to fall left, and likewise for 'fall right' parameters.
3. The blob is 90% likely to end up under the same chute as the ball.
4. The blob can't end up under the left chute if it travels right, and vice versa.

We will then use inequalities to establish bounds on $\mathbb{P}(D \mid A)$, i.e. the probabilities that the blob will travel left or right if the ball is released with 'fall left' or 'fall right' parameters.

$$
\begin{aligned}
\mathbb{P}(a_i, d_{m\neq i}) &= \sum_{j,k} \mathbb{P}(a_i, b_j, c_k, d_m) \\
&= \sum_{j} \mathbb{P}(a_i, b_j, c_m, d_m) \quad \text{since } \mathbb{P}(a_i, b_j, c_k, d_m) = 0 \text{ when } k \neq m \\
&= \mathbb{P}(a_i, b_i, c_m, d_m) + \mathbb{P}(a_i, b_m, c_m, d_m) \\
\mathbb{P}(a_i, d_{m\neq i}) &\leq 0.11 \quad \text{since } \mathbb{P}(b_i, c_{m\neq i}) = 0.1 \text{ and } \mathbb{P}(a_i, b_{m\neq i}) = 0.01 \\
\mathbb{P}(a_i) &= \mathbb{P}(a_i, d_i) + \mathbb{P}(a_i, d_{m\neq i}) \\
\mathbb{P}(a_i, d_i) &\geq 0.39 \quad \text{since } \mathbb{P}(a_i) = 0.5 \\
\mathbb{P}(d_i \mid a_i) &= \frac{\mathbb{P}(d_i, a_i)}{\mathbb{P}(a_i)} \geq 0.78
\end{aligned}
$$

In other words, it is at least 78% likely that the blob will travel left if the ball is released with 'fall left' parameters (and hence no more than 22% likely it will travel right if the ball is released with 'fall left' parameters). This proof relies on decomposing marginal probabilities into component terms, and the fact that probabilities are non-negative numbers which sum to one.

In contrast, to use Bayes' rule

$$\mathbb{P}(d_i \mid a_i) = \frac{\mathbb{P}(a_i \mid d_i)\mathbb{P}(d_i)}{\mathbb{P}(a_i)}$$

we would need to start by knowing $\mathbb{P}(a_i \mid d_i)$, the way in which the parameters of ball release statistically depend on the trajectory of the blob, as well as the statistics $\mathbb{P}(d_i)$ of the blob's trajectory (we already know that $\mathbb{P}(a_i) = 0.5$). We would then be able to deduce an exact number for $\mathbb{P}(d_i \mid a_i)$, in contrast with the inequality $\mathbb{P}(d_i \mid a_i) \geq 0.78$ that is imposed by the constraints actually given.

In general, for a pair of events a and b, knowing $\mathbb{P}(b)$ and $\mathbb{P}(b \mid \neg a)$ provides a meaningful lower bound on $\mathbb{P}(a)$, providing that $\mathbb{P}(b) > \mathbb{P}(b \mid \neg a)$.

$$\begin{aligned}\mathbb{P}(b) =& \mathbb{P}(a \wedge b) + \mathbb{P}(\neg a \wedge b) = \mathbb{P}(a \wedge b) + \mathbb{P}(b \mid \neg a)\mathbb{P}(\neg a) \\ \leq& \mathbb{P}(a) + \mathbb{P}(b \mid \neg a)\mathbb{P}(\neg a) \quad \text{since } \mathbb{P}(a \wedge b) \leq \mathbb{P}(a) \\ \leq& \mathbb{P}(a) + \mathbb{P}(b \mid \neg a)(1 - \mathbb{P}(a)) \\ \therefore \quad \mathbb{P}(a) \geq& \frac{\mathbb{P}(b) - \mathbb{P}(b \mid \neg a)}{1 - \mathbb{P}(b \mid \neg a)} \quad \text{by rearrangement.}\end{aligned}$$

Since $\mathbb{P}(a) \geq 0$ by definition, this constraint is meaningless unless $\mathbb{P}(b)$ is greater than $\mathbb{P}(b \mid \neg a)$.

References

Dennett, D. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Dennett, D. (2009). Intentional systems theory. In A. Beckermann, B. P. McLaughlin, & S. Walter (Eds.), *The Oxford handbook of philosophy of mind* (pp. 339–350). Oxford: Oxford University Press.

Fodor, J. A. (1978). Propositional attitudes. *Monist*, *61*(4), 501–523.

Friston, K. (2013). Life as we know it. *J R Soc Interface*, *10*(86), 20130475.

Horgan, T., & Woodward, J. (1985). Folk psychology is here to stay. *Philosoph Rev*, *94*(2), 197–226.

Maturana, H. R., & Varela, F. J. (1987). *The tree of knowledge: biological roots of human understanding*. Boston, MA: Shambhala.

Pearl, J. (2000). *Causality: models. Reasoning and inference*. New York: Cambridge University Press.

Rachlin, H. (1992). Teleological behaviorism. *Am Psychol*, *47*(11), 1371.

Scott, R. M., & Roby, E. (2015). Processing demands impact 3-year-olds' performance in a spontaneous-response task: New evidence for the processing-load account of early false-belief understanding. *PloS ONE*, *10*(11), e0142405.

Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind*. Cambridge, MA: Harvard University Press.